Lifetime Earnings Risk with Machine Learning

Ethan Ballou[*]

June 12, 2025

Abstract

Abstract. This is our abstract. It is abstract.

Keywords:
JEL Codes:

[*] University of Wisconsin - Milwaukee

1 Introduction

This is an example citation [?].

2 Literature Review

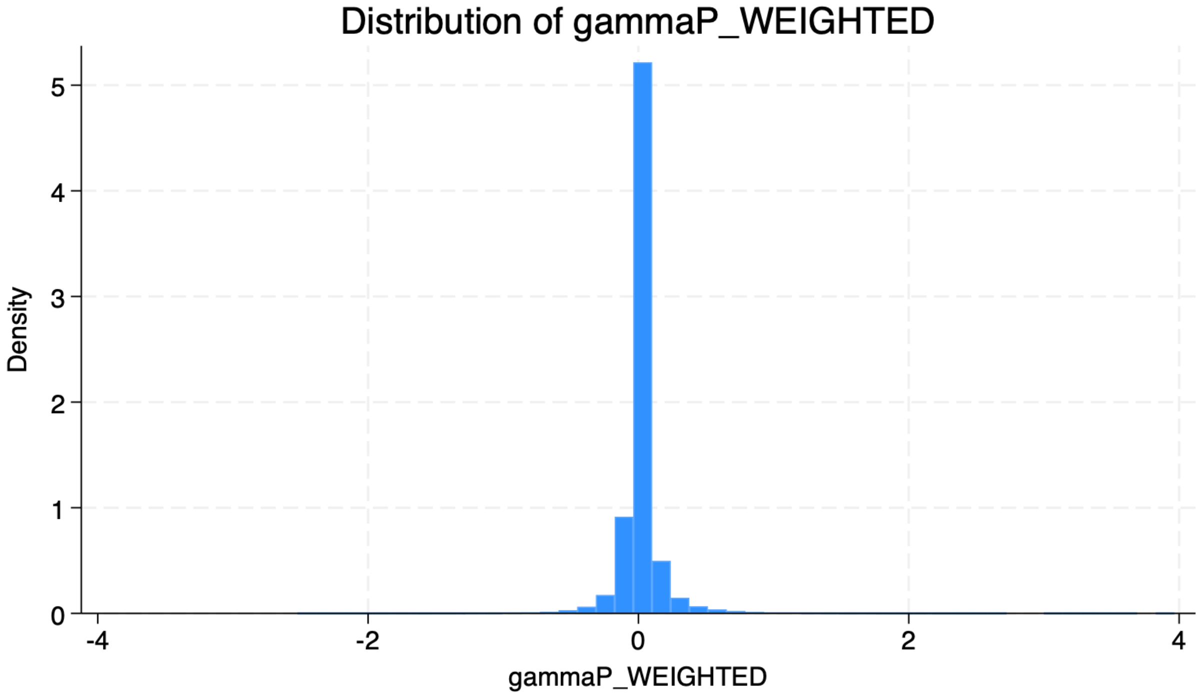

Figure 1: Distribution of gammaP_WEIGHTED

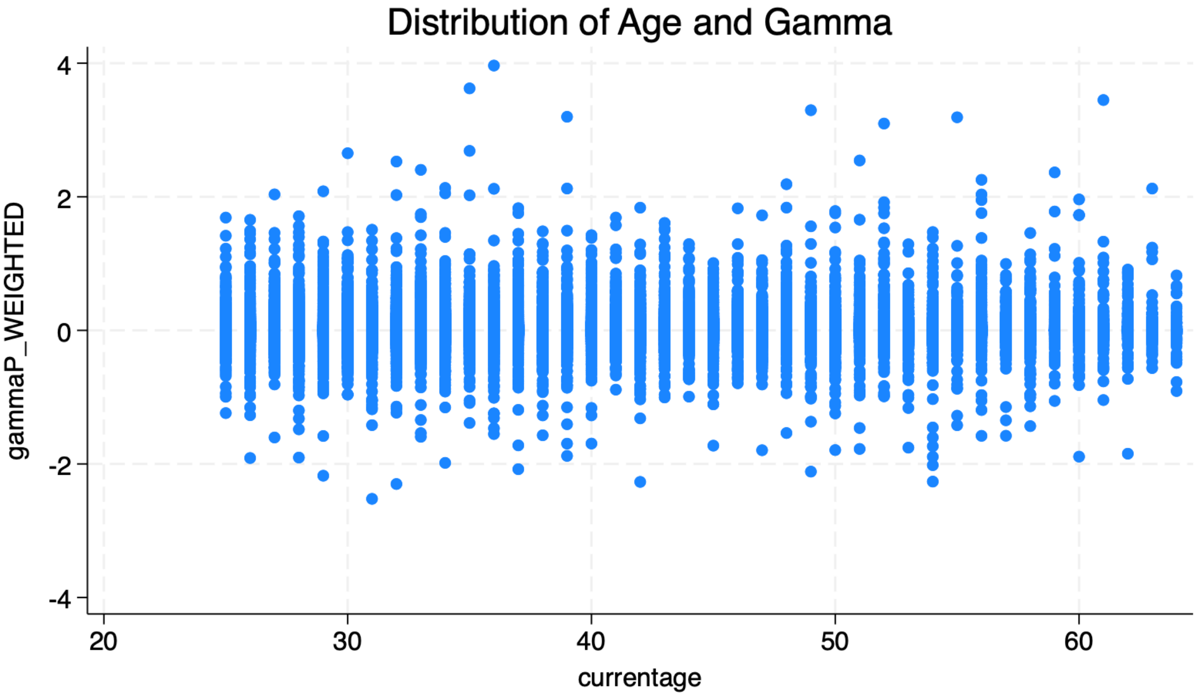

Figure 2: Scatterplot of Age vs. gammaP_WEIGHTED

3 Data and Model

start with model - Income process and equations - explain lightly RIP - Omega function - Then gamma function

- PSID data

4 Empirical Strategy

The empirical strategy of this paper is broken into two parts. The first part is estimating gamma and then constructing a weighted gamma for each individual in each year since gamma alone is across i, t, j, and q. (i - individual, t - time, j and q - period or ”window” used in calculation of gamma). The second part of the empirical strategy is estimating the then weighted gamma variable using various models and variables to look for trends in the characteristics of individuals in regards to lifetime earnings risk.

For the estimation of gamma and consolidation across j and q is done by using a mixed regression for the gamma and then a fixed effects regession to get the composit or weighted gamma for each person year.

- add mixed regression equation here - maybe FE regression?

The seond part of the empirical strategy uses various models to estimate lifetime earnings risk (gamma). The first model is a standard OLS regression of the lifetime earnings risk variable, gamma, on various controls and variables. The second model is a stepwise regression which selects variables based on a p-value threshold. The third model is a lasso regression which penalizes the size of the coefficients to select variables. Finally, the fourth model uses a multi-layer perceptron and SHAP values to interpret the results.

The OLS regressions are standard however the stepwise and lasso models do have componets that are worth mentioning. For the stepwise regression model a cutoff value of 0.05 is used to select variables. The model removes the least significant variable (or group of variables in the case of a set of controls) in rounds until it reaches the cutoff. 0.05 was selected so that the model would still select some variables but the rankings of the many of the variables would be clear. A higher cutoff and most of the variables would be selected and cardinal rankings would not be visible. Too low of a cutoff value and many of the variables would not be selected at all.

As for the lasso model, the model is set up to select variables based on the Bayesian Information Criterion (BIC) which is a common method for selecting variables in penalized regression. While the model selects lambda based on cross-validation, the selected model isn't very relevant to the analysis as the rankings are so the selected model isn't discussed.

Finally the multi-layer perceptron model is a neural network that is used to estimate lifetime earnings risk. The model is trained on the same variables and data used by the OLS, stepwise, and lasso models. The model is 4 dense layers (excluding the input layer) with 1000 nodes each and a linear output layer at the end. The model used a sigmoid activation function in the hidden layers to allow for continouous support. This was done instead of a ReLU activation function to allow for more definition in the parameters instead of some parts of the perception being ”dead” and not contributing and complicating the SHAP value interpretation. This model size and strucutre was selected as it achieved the best performance in terms of mean squared error (MSE). The model trained with MSE as its loss function and used early stopping to prevent overfitting and trained on 70 percent of the data with the rest being used for test and validation.

The SHAP values are then used to interpret the results of the multi-layer perceptron model. SHAP values are constructed by calculating marginal contribution of a variable as a deviation from the output variable's mean. This is done across a sample of observations and gives each variable a distribution of SHAP values. The SHAP values are then used to rank control variables based on the average SHAP for a given variable while the summary plot shows the distribution of SHAP values for the continuous variables.

5 Results

Gamma is a centered heavily around 0 with a standard deviation of 0.1686 as seen in Figure 1. Being centered at 0 is due to its derivation and as seen in Figure 2 there is not a clear correlation with age which might be expected. However there does seem to be a slight tightening past the age of 60 as the variance of gamma appears to decrease at least somewhat.

TABLE 1

The beginning of the analysis is just simple OLS regressions of gamma. The further analysis will focus on the which variables are most significant or important and less on the actual size of the effect. However size of coefficents is something OLS can easily address. While gamma is centered around zero, variables do still have effects despite being small. Table 1 shows the OLS estimates for gamma across different specifications.

The different specifications include different sets of controls, such as occupation and industry controls along wiht other controls such as state, year, cohort, and race.

The coefficents are quite small however some are larger than others. Less than high school and high school or some college education are lightly significant in some cases. They have some of the larger effects compared to many of the variables with less than high school being somewhare around -0.004 and -0.006 and highschool and some college being around -0.003 to -0.005.

The bachelors degree and less than a masters variable is smaller than the other two education variables which would imply college education does provide stabler employment and earnings overall. Howvever despite this the variable is not significant in any of the models suggests larger variation in earnings for those with a bachelor's degree or less than a master's degree. This variation could be due to variation in fields of study which would explain the larger coefficents in models 3, 4, and 5 where industry or occupation controls are included.

The other variables that are significant are the age variables. The age, age squared, and age cubed varaibles are all significant in all models. The interesting results is that the age squared and cubed variables are slightly more significant than the regular age variable. On some level it is expected that towards the end of a person's career they would become more risk averse and it is possible the squared and cubed terms capture this variation occuring right before retirement.

Overall the OLS results show that education and age are important in explaining lifetime earnings risk. This pattern contrinues in the stepwise and lasso results.

TABLE 2

Table 2 shows the stepwise results for the same 5 models. The stepwise results are based on a p-value threshold of 0.05. The stepsie regression results are able to assign cardinal rankings to how significant certain variables are in explaining lifetime earnings risk. In the stepwise regression the controls are treated as groups of variables such that the model cannot remove single variables from a set of controls and must remove the entire set of controls at once. The numbers assigned to the variables not selected indicate the the order in which they were removed usch that 1 is the last variable removed before the model is finalized.

The stewise models show similar results to OLS in that both education and age play an important role in predicting lifetime earnigns risk. Simlar to OLS less than highschool and high school and some college are the two education variables important to the model while the bachelors degree and less than a masters is not selected and not relevant in the model. And for the two models where the two important education variables aren't selected they are still the last two variables to be removed before the cutoff.

The other interesting variables that show up are the probability of recession and real GDP growth. These two variables are in the last 4 variables to be removed in all models suggesting their importance. Specifically probability of recession was the last variable removed in 3 of the models despite different controls. Probability of recession and real GDP growth were both important with and without industry controls which would be where one would expect at least some of their variation to be captured. This implies that variation across industry sensitivity to macroeconomic outcomes may not actually have much explanatory power in lifetime earnings risk.

Some surprising results regarding items not in the model is that industry controls are not selected by the stepwise regression in either of the models it is included in. This is despite the third model not even including occupation controls. Not only are industry controls not selected they are not close to the cutoff either, being the 7th and 10th variable from the cutoff in the two models. This can be contrasted with state controls which are slected in every model. The inclusion of state controls along with the importance of probability of recession and GDP growth may suggest that government policy plays an important role in lifetime earnings risk. However if this were the case industry level policy does not seem to be a part of the mechanism, as varaition in industry policy within states would be captured by the industry controls.

Finishing up the controls, occupation and state controls are selected in all models where they are included. This is in contrast to the rest of the controls being far from the cutoff and not seen as relevant in the analysis. The interpretation of state controls could a couple of things. The first was already mentioned in the previous paragraph regarding variation in government policy across states. However the second interpretation would be that the labor market enviroment is significnatly different across states and this is being captured.

While the labor market enviorment is heavily connected to governemnt policy, there are other factors that play a role. Labor markets are going to vary across states due to things like population density, resources, and other factors that are not directly related to government policy. In regard to occupation controls this is not very surprising as variation across job characteristics like qualifications, respoonsibilties and so forth is quite large even when considering within state and industry.

All in all the stepwise results show that education and age are important in explaining lifetime earnings risk. The stepwise results also show that macroeconomic variables such as probability of recession and real GDP growth are important and that the inclusion of state controls may further the hypothesis of government policy playing a large role. And finally the stepwise results show that occupation controls are important while industry controls are do not appear to play a large role.

TABLE 3

Now moving on to the lasso results shown in Table 3. As mentioned before lasso is a penalized regression and tries to force coeffiecnts to zero. This means lasso gives some weight to the size of the coefficents (relatively speaking) compared to stepwise which cores only about the p-value.

The first point to be made is that all the controls are selected in all cases. The configuration of this lasso model is the same as the stepwise model in that it can either indclude an entire set of controls or none of them.

Moving on from the controls, the two education variables that were important in the stepwise and OLS models are ranked quite high in importance as they were the in the top three variables in every case. This contiues the pattern of the less than high school and high school or some college variables playing an important role while the bachelor's degree and less than a master's variable is not as relevant.

However one pattern that does break here is the age variables. The age variables are among the least relevant variables in this analysis. Particularly the age squared variable which is often one of the last variables selected by the model. This is in contrast to the OLS results where the higher order age variables where slgihtly more significant than the regular age variable. The hypothesis there being that these higher order age terms are capturing a decrease in variation right before retirement. However in the lasso models this is not the case when it comes to the age squared variable. However it could be that this variation is better captured in the age cubed variable which is seem as more important than even the regualr age variable in 3 of the models.

The other observation to be made is regarding the probability of recession and real GDP growth variables. These two variables are not selected in any of the models except for the first model. This is in contrast to the stepwise results where they were both important.

SHAP HERE

Finally as summary of the shapley values are shown in Figure 3. This can add some detail to the analaysis that linear models cannot by representing the effects of a variable not in a scalar way but in a distributional way.

The table shows just the continous (or at least somewhat continuous) variables. Tenure, wage, and real GDP growth are shown to have relatively small effects regardless of the values. However the age variables display an interesting pattern. First the values are distributed such that lower values have a negative effect on gamma. However the more intersting observation is the skewness. All of the variables are skewed to the right as based on this for the really high values of age (which show up less often) the effect is positive. This is in contrast to the hypothesis that higher age may be related to a decrease in risk before retirement. However this is supported by the OLS results somewhat in that the cubed age variable was positive and very significant. That being said, based the table the age squared variable has the most variation in effects compared to the rest of the age variables which does still support that the result that the age becomes important towards the end of a person's career.

OCCUPATION AND INDUSTRY RESULTS

Throughout all the analysis occupation controls were included in every possible model. These occupations and industries are now further analyzed in Tables 4 and 5. The results are shown in a similar way to the previous tables by ranking the variables. The LASSO ranking is based on the order in which variables would enter the model if the penalty were relaxed (across values of lambda) and the SHAP ranking is based on the importance of the variable in predicting lifetime earnings risk.

TABLE 4

Starting with the analysis of occupations shown in Table 4,

6 Conclusion

Table 1: OLS Estimates for γ (Coefficients × 100)

	(1)	(2)	(3)	(4)	(5)
Education (less than 12)	-0.361	-0.360	-0.522^{*}	-0.597^{*}	-0.630^{**}
	(0.255)	(0.273)	(0.286)	(0.307)	(0.310)
Education (12 to 14)	-0.283	-0.315^{*}	-0.418^{**}	-0.417^{*}	-0.455^{**}
	(0.175)	(0.181)	(0.193)	(0.216)	(0.219)
Education (14 to 16)	-0.089	-0.063	-0.138	-0.114	-0.147
	(0.206)	(0.209)	(0.216)	(0.228)	(0.230)
PrRecess	-0.006	-0.038	-0.037	-0.040	-0.039
	(0.005)	(0.036)	(0.036)	(0.036)	(0.036)
rGDPgrow	-0.033	0.076	0.071	0.063	0.058
	(0.033)	(0.172)	(0.172)	(0.172)	(0.172)
fhwage0_P0	-0.006	0.004	0.005	0.002	0.004
	(0.025)	(0.027)	(0.027)	(0.028)	(0.028)
ma5aep	0.004	0.003	0.004	0.004	0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
veteran	0.021	-0.003	0.053	0.018	0.040
	(0.141)	(0.151)	(0.153)	(0.154)	(0.155)
OLF	0.627	0.584	0.550	0.619	0.625
	(0.627)	(0.628)	(0.628)	(0.629)	(0.629)
tenure	-0.010	-0.009	-0.008	-0.012	-0.010
	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
currentage	0.874^{**}	0.936^{**}	0.926^{**}	0.919^{**}	0.907^{**}
	(0.380)	(0.384)	(0.385)	(0.385)	(0.385)
currentagesq	-0.023^{**}	-0.025^{***}	-0.025^{***}	-0.025^{***}	-0.024^{***}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
currentagecube	0.0002^{***}	0.0002^{***}	0.0002^{***}	0.0002^{***}	0.0002^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Occupation Controls				✓	✓
Industry Controls			✓		✓
Other Controls		✓	✓	✓	✓

Notes: Standard errors in parentheses. Other controls include state, year, race, and cohort fixed effects. Statistical significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. All coefficients and standard errors are multiplied by 100 for easier interpretation.

Table 2: Stepwise Results for γ

	(1)	(2)	(3)	(4)	(5)
EDU1	selected	2	2	selected	selected
EDU2	selected	1	1	selected	selected
EDU3	6	10	8	6	6
PrRecess	1	3	3	1	1
rGDPgrow	4	4	4	4	4
fhwage0_P0	7	11	13	9	9
ma5aep	2	6	6	3	3
veteran	8	12	12	10	11
OLF	3	5	5	2	2
tenure	5	7	9	5	5
currentage	selected	selected	selected	selected	selected
currentagesq	selected	selected	selected	selected	selected
currentagecube	selected	selected	selected	selected	selected
Occupation Controls	-	-	-	selected	selected
Industry Controls	-	-	7	-	10
Cohort Controls	-	8	10	7	7
Race Controls	-	13	14	11	12
Year Controls	-	9	11	8	8
State Controls	-	selected	selected	selected	selected
Occupation Controls				✓	✓
Industry Controls			✓		✓
Other Controls		✓	✓	✓	✓

Notes: This table reports results from stepwise regression models using a p-value threshold of 0.05. ”Selected” indicates variables retained in the final model. Numbers indicate the order of variable removal (with 1 being the last variable removed before model finalization). ”-” indicates the variable was not included in the initial model specification.

Table 3: Lasso Results for γ

	(1)	(2)	(3)	(4)	(5)
EDU1	3	3	2	1	1
EDU2	2	1	1	1	1
EDU3	8	7	6	6	4
PrRecess	4	Not Selected	Not Selected	Not Selected	Not Selected
rGDPgrow	6	Not Selected	Not Selected	Not Selected	Not Selected
fhwage0_P0	7	9	9	8	6
ma5aep	1	2	1	2	1
veteran	10	8	8	7	5
OLF	3	2	3	1	1
tenure	5	4	4	3	2
currentage	9	6	7	5	4
currentagesq	11	10	10	8	7
currentagecube	2	5	5	4	3
Occupation Controls	-	-	-	Selected	Selected
Industry Controls	-	-	Selected	-	Selected
Cohort Controls	-	Selected	Selected	Selected	Selected
Race Controls	-	Selected	Selected	Selected	Selected
Year Controls	-	Selected	Selected	Selected	Selected
State Controls	-	Selected	Selected	Selected	Selected
Occupation Controls				✓	✓
Industry Controls			✓		✓
Other Controls		✓	✓	✓	✓

Notes: This table reports variables selected by Lasso regression with Bayesian Information Criterion (BIC) variable selection. "Selected" indicates variables retained in the final model. Numbers in parentheses indicate the order in which variables were added to the model across variation in lambda. "-" indicates the variable was not included. "Not Selected" indicates the variable was not selected by Lasso for any lambda used in cross validation but was provided in the model specification.

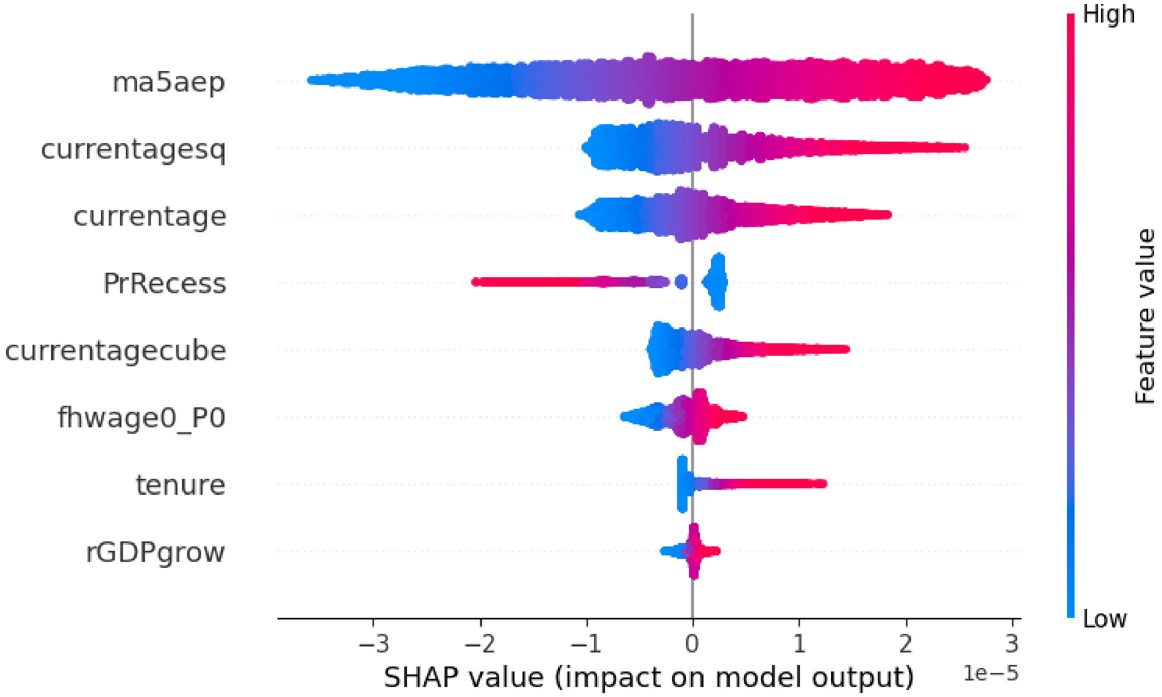

Figure 3: SHAP Summary Plot

Table 4: Lasso and SHAP Results for Occupations

Occupation	SHAP Rank	LASSO Rank	Occupation	SHAP Rank	LASSO Rank
occ_21	1	Not Selected	occ_60	2	16
occ_84	3	26	occ_61	4	13
occ_1	5	14	occ_45	6	20
occ_70	7	Not Selected	occ_98	8	14
occ_37	9	Not Selected	occ_97	10	27
occ_2	11	20	occ_99	12	22
occ_13	13	14	occ_9	14	24
occ_11	15	23	occ_83	16	21
occ_4	17	3	occ_101	18	Not Selected
occ_95	19	25	occ_79	20	15
occ_85	21	Not Selected	occ_6	22	8
occ_999	23	Not Selected	occ_8	24	Not Selected
occ_93	25	17	occ_55	26	21
occ_20	27	16	occ_53	28	23
occ_17	29	6	occ_19	30	Not Selected
occ_5	31	18	occ_40	32	23
occ_58	33	26	occ_34	34	19
occ_59	35	13	occ_7	36	Not Selected
occ_77	37	21	occ_87	38	21
occ_50	39	19	occ_14	40	Not Selected
occ_54	41	28	occ_96	42	15
occ_38	43	20	occ_3	44	12
occ_15	45	14	occ_32	46	14
occ_18	47	11	occ_42	48	11
occ_73	49	Not Selected	occ_31	50	Not Selected
occ_62	51	21	occ_44	52	17
occ_33	53	19	occ_63	54	28
occ_49	55	18	occ_86	56	17
occ_36	57	Not Selected	occ_43	58	16
occ_39	59	Not Selected	occ_74	60	17
occ_72	61	Not Selected	occ_64	62	2
occ_56	63	Not Selected	occ_92	64	Not Selected
occ_80	65	Not Selected	occ_88	66	13
occ_12	67	14	occ_75	68	25
occ_81	69	Not Selected	occ_30	70	Not Selected
occ_35	71	9	occ_57	72	Not Selected
occ_89	73	Not Selected	occ_71	74	Not Selected
occ_16	75	29	occ_94	76	25
occ_82	77	Not Selected	occ_52	78	Not Selected

Notes: This table reports occupations selected by Lasso regression with Bayesian Information Criterion (BIC) for predicting earnings risk. "SHAP Rank" shows the variable importance ranking based on SHAP values (lower numbers indicate greater importance). "LASSO Order" indicates the order in which variables would enter the model if the penalty were relaxed. Note that the BIC-optimal model contained no occupation variables.

Table 5: Lasso and SHAP Results for Industries

Industry	SHAP Rank	LASSO Rank
twoind_3	1	28
twoind_9	2	10
twoind_19	3	18
twoind_30	4	2
twoind_16	5	19
twoind_21	6	8
twoind_14	7	Not Selected
twoind_18	8	Not Selected
twoind_5	9	7
twoind_33	10	3
twoind_10	11	21
twoind_4	12	27
twoind_29	13	21
twoind_999	14	17
twoind_15	15	16
twoind_7	16	13
twoind_11	17	16
twoind_12	18	3
twoind_22	19	27
twoind_1	20	11
twoind_25	21	10
twoind_27	22	3
twoind_6	23	23
twoind_20	24	11
twoind_23	25	18
twoind_8	26	28
twoind_24	27	29
twoind_31	28	23
twoind_28	29	7
twoind_13	30	25

Notes: This table reports industries selected by Lasso regression with Bayesian Information Criterion (BIC) for predicting earnings risk. ”LASSO Selection Order” indicates the order in which variables would enter the model if the penalty were relaxed. ”SHAP Ranking” shows the variable importance ranking based on SHAP values (lower numbers indicate greater importance).

References